AD-A042 966 MEMPHIS STATE UNIV TENN F/G 9/5
FLEXILEVEL ADAPTIVE TESTING PARADIGM: HIERARCHICAL CONCEPT STRU--ETC(U)
JUL 77 D N HANSEN, S ROSS, D A HARRIS F41609-75-C-0040
UNCLASSIFIED AFHRL-TR-77-35(II) NL

1 OF 1
ADA042 966

END
DATE FILMED
9-77
DDC

AFHRL-TR-77-35(II)

AIR FORCE HUMAN RESOURCES LABORATORY

FLEXILEVEL ADAPTIVE TESTING PARADIGM: HIERARCHICAL CONCEPT STRUCTURES

By

Duncan N. Hansen
Steven Ross
Memphis State University
Memphis, Tennessee 38152

Dickie A. Harris, Capt, USAF

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado 80230

July 1977
Final Report for Period May 1975 – March 1977

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Memphis State University, Memphis, Tennessee 38152, under contract F41609-75-C-0040, project 1121, with Technical Training Division, Air Force Human Resources Laboratory (AFSC), Lowry Air Force Base, Colorado 80230. Captain D. A. Harris, Instructional Technology Branch, was the contract monitor.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

MARTY R. ROCKWAY, Technical Director
Technical Training Division

DAN D. FULGHAM, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFHRL-TR-77-35(II)	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)* FLEXILEVEL ADAPTIVE TESTING PARADIGM: HIERARCHICAL CONCEPT STRUCTURES.	5. TYPE OF REPORT & PERIOD COVERED Final rept. May 1975 – March 1977
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Duncan N. Hansen Steven Ross Dickie A. Harris	8. CONTRACT OR GRANT NUMBER(s) F41609-75-C-0040 new
9. PERFORMING ORGANIZATION NAME AND ADDRESS Memphis State University Memphis, Tennessee 38152	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62703F 11210309 12 13
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235	12. REPORT DATE July 1977 13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* Technical Training Division Air Force Human Resources Laboratory Lowry Air Force Base, Colorado 80230	15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

adaptive model
adaptive testing
computer-assisted testing
hierarchical concept structures
technical training

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Recent research indicated the benefits of computerized adaptive testing for assessing achievement in technical training. In a prior study ~~(Hansen, Harris, & Ross, 1976)~~, results indicated that, relative to a conventional test, the adaptive strategy significantly reduced testing time while yielding equivalent parametric outcomes. The present study extended this research by examining the feasibility of a similar model applied over a hierarchically arranged series of subtests in a more sophisticated instructional context. As in the initial study, the adaptive model was a modification of Lord's flexilevel algorithm which allows students to move systematically among easier and harder items according to a response contingent rule; an individualized entry component, however, was not employed in this particular application. The course selected for evaluating the model was the Precision Measurement Equipment

→ next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/C 390023

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

Item 20 (Continued):

cont

Specialist Course taught at Lowry Air Force Base, CO. A total of 133 students participated in the study in fulfillment of achievement testing requirements for Blocks II and IV of the course. The two Block achievement tests were each divided into five hierarchically related subtests so as to allow for the assessment of sequential performance contingencies. Data collection involved a within-subject design in which students were entered at the median of the initial subtest and administered items by the flexilevel procedure. Following completion of the adaptive test, all remaining items were administered. The same procedures were then followed for the remaining subtests in the hierarchy. Test validity analyses yielded part-whole correlations between adaptive test and total test scores (r's = .95). Descriptive test indices and test reliabilities were also essentially identical. Importantly, the time savings associated with adaptive testing approximated 30 percent for both blocks. Additional findings concerned the interrelations among subtest outcomes both within and between blocks. The results were interpreted as clearly supporting the generalizability of adaptive testing benefits to highly complex, hierarchically structured training.

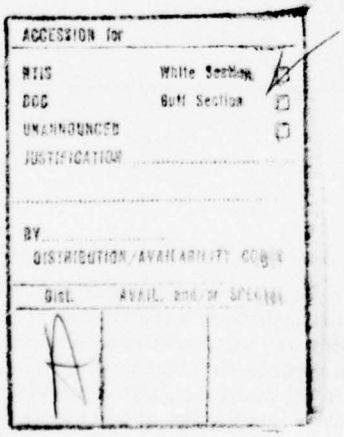


TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

FLEXILEVEL ADAPTIVE TESTING PARADIGM: HIERARCHICAL CONCEPT STRUCTURES

Introduction

Technical training spans a broad range of scientific instruction. At the lower end of the continuum, one finds systems-oriented procedural instruction for such career patterns as clerks and warehouse personnel. At the upper end there are technical training courses addressed to the maintenance and standardization of the test equipment itself. Over this broad continuum, technical training has continued to evidence a need for refinement of its measurement processes. There have been two primary reasons for this extended investigation of measurement within a technical training environment. First, a commitment to skill mastery and on-the-job competencies had to be based on increased course test accuracy and validity. In turn, the amount of time given to testing within a training course brought forth a need for reductions without sacrificing psychometric properties. Adaptive testing, a process by which only selective items are presented to a given student, offers promise both in maintaining psychometric properties of the test and in yielding significant time savings. The purpose of this study was to generalize earlier findings on adaptive testing (Hansen et al., 1976) by applying it to a highly sophisticated, hierarchically arranged technical training course.

Computer-based adaptive testing (CAT) paradigms provide a process by which optimal items are presented which assess a student's critical performance. The adaptive process attempts to remove items which are either too easy or too difficult. A recently completed study on a less technical course in Air Force Inventory Management indicated that the reliability and validity coefficients for adaptive tests were not only essentially equivalent to those of conventional tests but also yielded an approximate 40 percent time saving (Hansen et al., 1976). These time savings were similar to those reported by Waters (1975) and Larkin and Weiss (1975). However, in neither of these empirical studies was a highly sophisticated course selected, especially one with hierarchical concept and skill structures.

Training courses based on high levels of technology typically involve a complex hierarchy of concepts and skills. Based on empirical task analysis and Gagne's (1962) theory of hierarchies of learning, the materials of these technical courses can be represented by tree structures having subordinate-supraordinate relationships. The challenge for measurement is to identify students having deficiencies in critical concepts while assessing mastery of all elements of the conceptual structure. For adaptive testing the challenge is to prove its feasibility within this

training context and establish its level of reliability and validity in detecting subconcept or skill deficiencies. Adaptive testing can be used to establish mastery levels within the hierarchy as demonstrated by Ferguson (1969) via the use of iterative movement over levels. This feature of adaptive testing was not employed in this study because of the desire to assess the hierarchical relationships from the subordinate to the supraordinate, that is, to assess the relationship among levels in the course hierarchy.

The purpose of the study was to evaluate the feasibility of CAT. As a setting for this feasibility assessment of adaptive testing, the Precision Measurement Equipment Specialist Course (PME) located at Lowry AFB Technical Training Center, Colorado was selected. Mastery tests were imbedded within two beginning two-week instructional blocks. Feasibility was to be judged in terms of adaptation of the students to the terminal as well as in terms of the operational characteristics. Most importantly, the relationship of the adaptive test score with the conventional test scores was to be assessed. A within-subject design which estimated concurrent validity was utilized. The flexilevel item selection algorithm developed by Lord (1971) was employed. This allowed a student to move systematically among harder and easier items according to a response contingent rule. The primary purpose of the study was to assess the psychometric outcomes as well as time savings offered by adaptive testing of hierarchical concepts. Secondary questions dealt with the hierarchical relationships among the identified subtest sections so as to further assess the degree to which students could branch from concept to concept or possibly omit concepts.

Method

The focus of the study was to assess the feasibility and validity of flexilevel testing of a series of hierarchically arranged subconcepts. Procedures for data collection involved the use of a within-subject design in which students were entered via a computer terminal sign-on process at the median item difficulty level of the first subtest and then administered items by means of flexilevel adaptive movement procedures. After the student completed the adaptive portion of the subconcept test, all remaining items were presented. The student then entered the next higher subtest. Thus, both an adaptive score and a conventional test score for each subtest were obtained for each subject in the sample. The within-subject design refers to the multiple repeated measurement of students through subtests.

The major independent variables consisted of: a) the testing strategy--adaptive and conventional and b) a replication on two different units of instruction (Blocks II and IV). Dependent variables consisted of: a) conventional test scores, b) flexi-scores, c) number of flexi-test items, d) flexi-time, e) total test time, and f) errors after flexi-exit. Reliability estimates of all test forms were obtained by means of the KR-20 procedure and linear combination of subtests (Nunnally and Durham, 1975) at both item and form levels.

Subjects

The subjects consisted of 133 enlisted personnel enrolled in the PME Course. The student population, representing the three services, from which the subjects were selected was slightly variable in characteristics pertaining to age, educational background, career goals, and military experience. Student characteristic data collected during the past year indicates that the typical enrollee in the *Precision Measurement Equipment* Specialist Course is male (87.5 percent male to 12.5 percent female), an average age of 25 (S.D. = 5.66), varied in educational background (pre-high school - 54 percent; high school - 36 percent; and collegiate - 10 percent), and varied in military experience. This course is a Tri-Service Activity (Air Force - 60 percent; Army - 12 percent; Marine - 10 percent; and civilian/foreign - 18 percent).

Subjects were oriented to believe that participation in the study simply involved taking their regularly assigned achievement test (Block II o IV) under a newly developed computer-assisted test administration system, that is, at an interactive computer terminal. Since the transition from adaptive to conventional item presentations took place without interruption or change in normal test-taking procedures, it was considered doubtful that subjects became aware of the purposes of the experiment, or for that matter, that they even suspected that there was anything unusual about the selection or sequencing of items as compared to the conventional paper-and-pencil procedures. Since test question-answer review was a part of the instructional approach (e.g., "Do the easy questions first, don't stay too long on the difficult ones."), the computer program allowed for post test response paging (i.e., student control of item representation) and answer changing. For this study, only the subjects' first responses to items were considered.

Hierarchical Test Structure. The Block II and IV Tests of the PME Course were selected for use in validating the adaptive

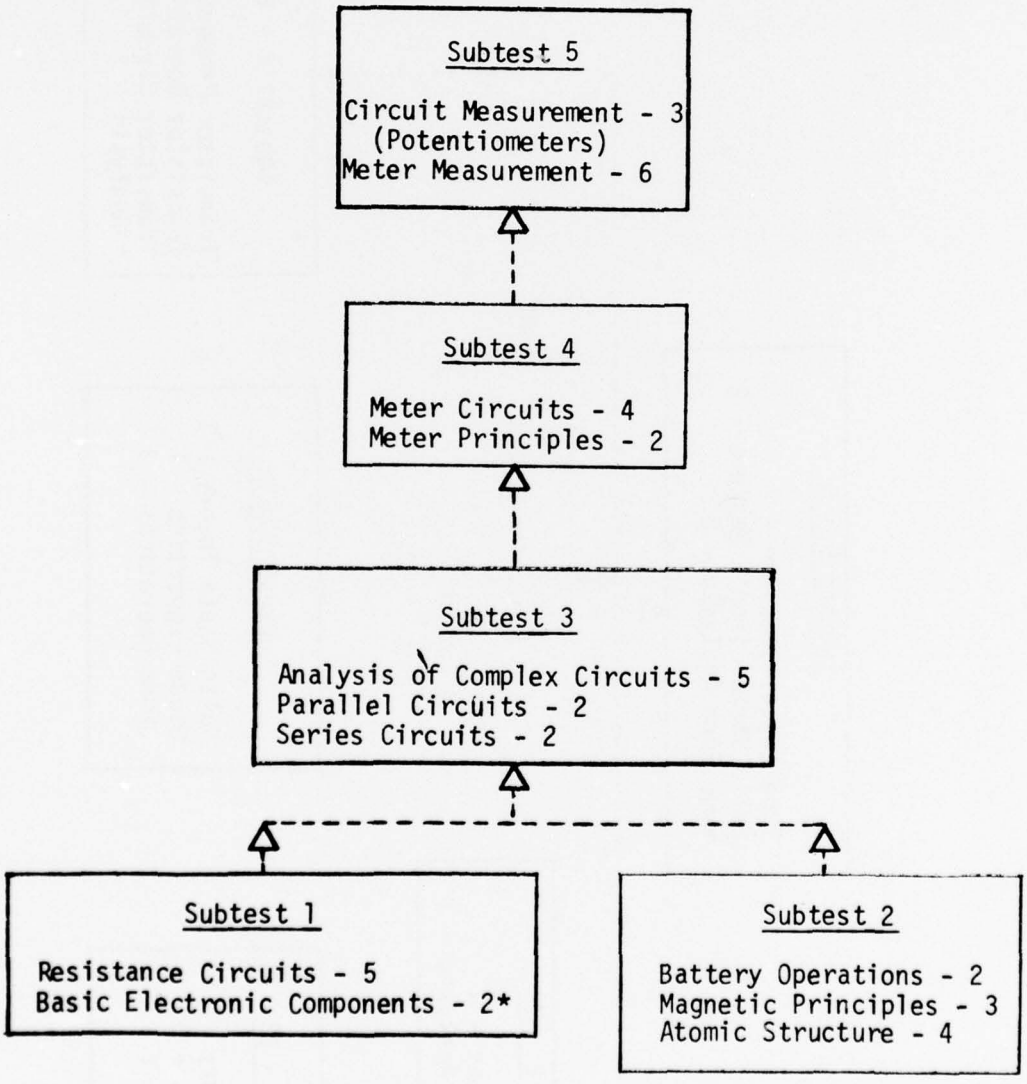
testing paradigm since each possessed obvious hierarchical structuring of concepts. Satisfactory performance on each test was prerequisite to further progress in the course. Each block test consisted of 40 multiple-choice items, each containing four response alternatives. Graphic illustrations were presented within a printed test booklet; test items were presented via the computer terminal.

The hierarchical conceptual structure of the Block II Test is presented in Figure 1. Subtest 1 (Basic Electrical Components) and Subtest 2 (Magnetic Principles) form the parallel inputs into a linear tree structure. This hierarchical structure was relatively simple, but allowed for the assessment of sequence contingencies (e.g., to what degree did test performance on a given level predict higher order performance?). Subtests were formed based on a task analysis of test items so as to represent homogeneous concepts, identifiable components in the concept tree structure, and sufficient numbers of test items to allow for flexilevel assessment. The numbers to the right of the concept label (See Figures 1 and 2) indicate the number of test items.

The hierarchical conceptual structure of the Block IV Test is presented in Figure 2. Subtests 1 and 2 (Electron Theory and Tube Operations) were sequentially related, while Subtests 3 and 4 were two parallel paths into the final culminating Subtest (5) on Wave Form Analysis. Thus, the nature of the subordinate-supraordinate relationships were variable and allowed for assessment of the task structure due to the two different pathways.

Procedure. Preparation activities involved frequent meetings with course instructors and supervisory personnel from PME two months prior to the conduct of the actual study. (Terminal installation and test item layout required numerous revisions.) The purpose of these meetings was to insure that the teaching staff understood the procedures they would be required to follow in coordinating the test administration and data collection. This was accomplished mostly through discussion and demonstration activities. Additionally, all instructors received a manual which provided a brief overview of the purposes of adaptive testing along with a detailed step-by-step account of the operational requirements for the present flexilevel test; that is, procedures for "signing on" the system, responding to items, interpreting and recording results, and "signing off."

The PME Course followed a criterion-referenced format in which testing occurred once per week. While each Block (II and IV) was two weeks in duration, multiple shifts offered about ten



* Number of test items

Figure 1. Hierarchical Structure for Total Block Test II

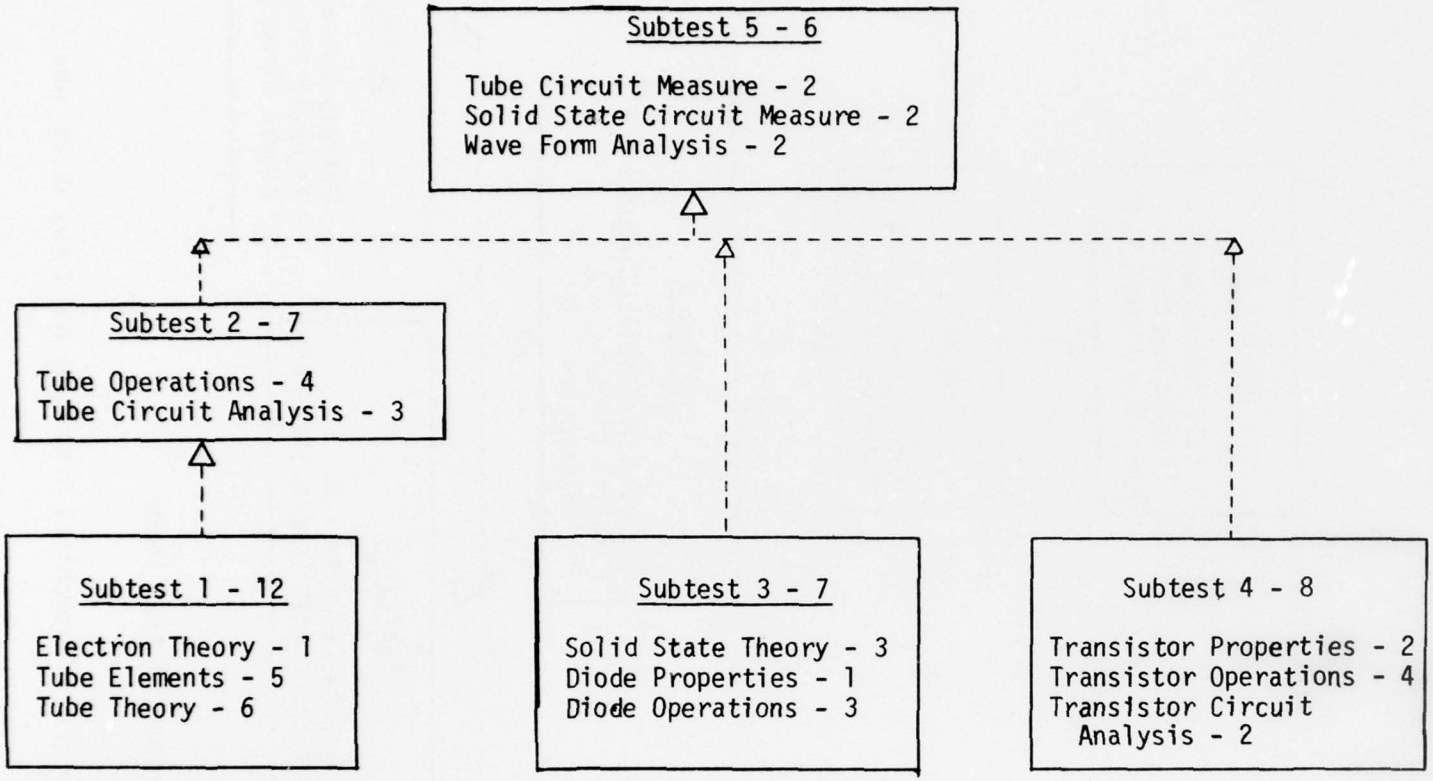


Figure 2. Hierarchical Structure for Block Test IV

new students per week. The adaptive test was administered on an individual basis, according to an instructor-controlled schedule. On testing days the students were directed to the computer terminal (a PLATO plasma screen terminal connected to the University of Illinois) and given instructions for "signing on." Various panel displays then appeared in a prearranged sequence with progression from one display to another dependent upon the student keypunching appropriate symbols or words.

After "signing on", students were requested to enter their names and Social Security numbers. Students unfamiliar with the system were then given instructions for taking the computer-based test and for using the PLATO system in general. These instructions could be recalled any time questions arose during the actual test; also, students were encouraged to seek assistance from their instructor if ever uncertain about the proper procedures for responding.

Following these preliminary instructions, students were entered into the flexilevel test at the median difficulty item for the first subtest. Test items were administered sequentially with the rate of presentation determined entirely by the student. Procedures for responding simply involved keypunching the numbers of selected multiple-choice alternatives. Students were told to carefully consider their responses before continuing with the next item. If dissatisfied with their initial choice, they were to erase it and select another alternative; if satisfied, they were to finalize their answer by requesting that a new item be presented. Once answers were finalized, they could no longer be changed except during a post-test item review process. This process was repeated for each subtest.

For the flexilevel portion of the test, the sequencing of items was determined in the following manner: once students were entered in the test at the median difficulty item, they were moved up and down the difficulty hierarchy based upon their performance. Specifically, each wrong response resulted in the presentation of the next easier item whereas each correct response resulted in the presentation of the next harder item. Unlike Lord (1971) who used fixed item length cutoff, in this study the cutoff was the completion of either the easiest or hardest item (ends of the test). After exiting out of the subtest at either the top or bottom level, they were administered all remaining items. The next subtest was then presented. At the completion of the entire test, the instructor was called to the terminal where he was able to obtain a summary of the student's performance. The specific information provided consisted of: a) total test scores, b) individual item scores,

c) total test time per subtest; and for use by members of the research team, d) predicted score, e) flexilevel exit, and f) flexilevel test time. A printed copy of the data was typically made available on the following day.

Computer Operation. Figure 3 presents a flow chart of a student moving through each of the steps. In signing on, the student entered his name and the computer executed a security check designed to limit system accessibility and assure test security. The computer system entered the student into an appropriate flexilevel test at the median. Thus, all tests were individually tailored to the student's current status.

Results

Table 1 presents the descriptive statistics for Blocks II and IV terminal tests of the PME Course. For each of the 40-item tests segmented into five hierarchical subtests, the total means and standard deviations were statistically nonsignificant ($p>.05$). As a variant scoring method, the Green scoring procedure (Green, 1970), in which the item difficulties of correct items only are averaged, yielded similarly patterned results ($p>.05$). Finally, the mean number of errors after the cutoff was less than one ($\bar{X}_2 = .80$ and $\bar{X}_4 = .44$ points, respectively). This might be attributed to the few items remaining after cutoff.

In terms of testing time, results of Blocks II and IV had essentially the same magnitudes. The time savings for the adaptive paradigm were 30 percent and 25 percent, respectively, for Blocks II and IV. As noted above, there were no significant differences between Blocks II and IV in terms of either performance or time. A review of the skewness indices indicates that normality was being approximated. In turn, the kurtosis indices were less than one and positive. This is important to both the significance testing and the reliability estimates.

Table 2 presents the descriptive statistics for the post-tests from both Blocks II and IV. If one considers subtests of equal length (lengths of six and seven items), the six-item subtests means (Block II--four and Block IV--five) were of similar magnitude while the seven-item subtests (Block II--one and Block IV--two, three) were more variable. These comparisons were confounded by position in the test hierarchy or implied level of conceptual difficulty.

A combination of mean exit item and subtest time allowed for

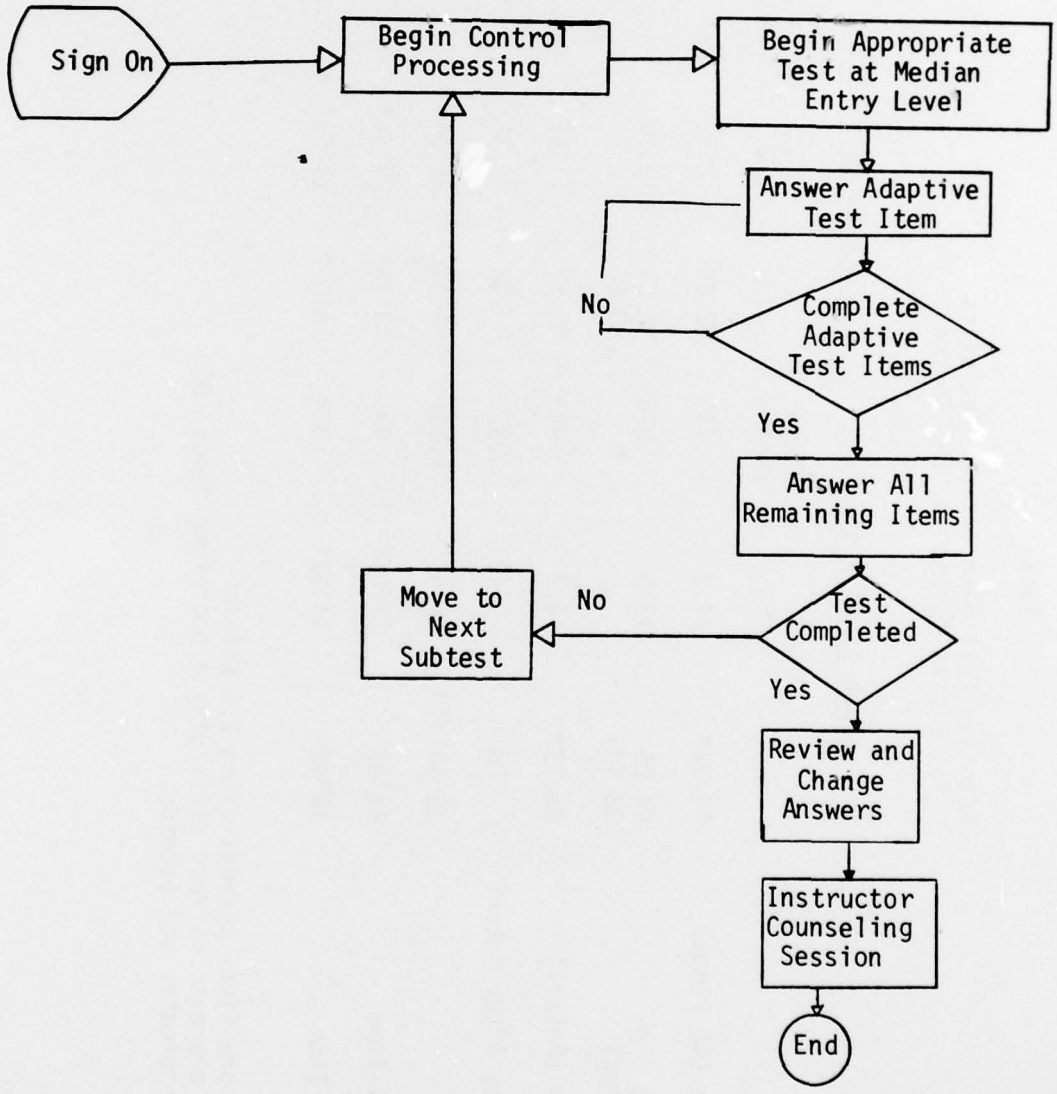


Figure 3. Flowchart of Student Progress Through Flexilevel Testing Program

Table 1

Adaptive Test Descriptive Statistics
for Blocks II and IV, PME Course

Variable	Block II (N = 61) $\bar{X}$	SD	Skew Index	Block IV (N = 72) $\bar{X}$	SD	Skew Index
Total Score (40 items)	84.34*	14.8	- .919	82.06	12.32	-1.00
Flexi-Score (Mean Items)	84.05 (28.43)	13.6	- .826	81.57 (30.83)	12.14	- .962
Green Score (Points)	85.13**	14.7	- .023	83.57	11.65	- .210
Total Errors After Cutoff	.80	.42	.872	.44	.32	.593
Total Time	65'45"***	39'41"	1.131	64'37"	36'04"	.988
Total Flexi-Time	46'05"	27'05"	.931	48'30"	29'51"	.903
Total Post Time	19'40"	14'37"	.874	16'07"	11'58"	.901

* Scores are items correct times 2.5 points.
** Scores are mean correct difficulty proportions times 100.
*** Time in minutes and seconds.

Table 2

Descriptive Performance and Time Measures
for Each Subtest in Block II and IV

Variable	Block II $\bar{X}$	Block II S.D.	Block IV $\bar{X}$	Block IV S.D.
Subtest 1	(7 items)		(12 items)	
Exit Item	4.56	.83	9.15	1.65
Post Items	2.44	.83	2.85	1.61
Errors After Cutoff	.03	.18	.01	.12
Adapt Time	7'51"*	9'05"	13'34"	15'01"
Post Time	4'13"	5'02"	4'57"	5'30"
Subtest 2	(9 items		(7 items)	
Exit Item	6.07	1.18	5.47	1.00
Post Items	2.93	1.18	1.53	1.00
Errors After Cutoff	.05	.22	.01	.12
Adapt Time	9'56"	10'32"	8'05"	8'09"
Post Time	3'57"	4'43"	2'55"	1'01"
Subtest 3	(9 items)		(7 items)	
Exit Item	6.08	1.33	4.72	.90
Post Items	2.92	1.32	2.28	.87
Errors After Cutoff	.05	.22	.03	.17
Adapt Time	9'48"	7'04"	7'51"	6'27"
Post Time	4'46"	4'02"	3'13"	4'07"
Subtest 4	(6 items)		(8 items)	
Exit Item	4.92	.88	6.68	1.10
Post Items	1.08	.87	1.32	1.07
Errors After Cutoff	.05	.22	.03	.17
Adapt Time	8'15"	4'37"	11'56"	10'46"
Post Time	2'39"	1'42"	2'07"	4'28"
Subtest 5	(9 items)		(6 items)	
Exit Item	6.80	1.48	4.81	.76
Post Items	2.20	1.46	1.19	.72
Errors After Cutoff	.02	.13	.03	.16
Adapt Time	10'15"	5'52"	7'04"	8'02"
Post Time	3'55"	1'58"	2'55"	4'03"

* Time in minutes and seconds

a comparison of mean time per item. For the Block II test, there were an average of 11.57 items after adaptive cutoff, or a 28.93 percent item saving. In turn, there was a 30 percent time saving between adaptive and total testing. The mean item time was 1.62 minutes for adaptive items while the mean post cutoff item time was 1.70 minutes. For the Block IV test, there were 9.17 items after cutoff which allowed for a 22.93 percent item saving. The time saving was 24.92 percent. The adaptive mean item time was 1.57 minutes while the post item time was 1.76 minutes. Item time calculations by subtest indicated that in six of the ten subtests the mean adaptive item time was less than the post item time, a finding somewhat counter to item times for adaptive ability testing. Examination of the post item times indicated that the poorer performing students who exited at the easy end of the subtest had excessively longer post item times on the highly difficult items.

In reference to the psychometric outcomes, Table 3 presents the mean item difficulties by subtest plus a Kuder-Richardson reliability index. A review of the mean item difficulties indicated that there was a progressive reduction in performance as task difficulty increased. The KR-20 for the Block II test was .841; the Block IV test had a coefficient of .788. Utilizing a mean item cutoff for each of the subtests that would retain at least 50 percent of the subject population, an adaptive Kuder-Richardson index was calculated. This was found to be r = .701 for Block II test and r = .714 for Block IV test. Assessing the effect of a full 40 items by the use of the Spearman-Brown Formula, the adaptive coefficients increased to Block II: r = .799 and Block IV: r = .782. For criterion tests, the KR-20 coefficient may be an underestimate due to the nonnormality of the item distribution. Haladyna (1974) reported a series of empirical studies that found the magnitude of the underestimation was slight in nature. Fortunately the separate subtest reliabilities can be combined linearly (Nunnally, 1967). Table 3 presents the linear combination reliability coefficients for total scores and adapt scores. An inspection of these coefficients indicated that only a slight difference occurred.

As to an expected hierarchical progression, the matching between the mean item difficulties in Table 3 and the hierarchical structures in Figures 1 and 2 presents some inconsistencies. For example, Block II was a relatively linear structure as based on the task analysis; the performance consistently dropped until the final subtest (5) on Circuit Measurement. This increase in final node performance was not expected by hierarchical learning theory. In turn, Block IV was a more complex hierarchical structure and again yielded a higher performance value on the final culminating

Table 3

Mean Item Statistics and Reliability Indices

	Block II (N = 61)		Block IV (N = 72)	
Subtest	Item N	Mean Item Difficulty	Item N	Mean Item Difficulty
1	7	.918	12	.810
2	9	.888	7	.776
3	9	.872	7	.895
4	6	.738	8	.762
5	9	.781	6	.838
Total KR-20		.841		.788
S.E. Measurement		2.03		2.208
Adapt KR-20		.701 (28 items)		.714 (31 items)
S.E. Measurement		1.57		1.749
Linear Combination Reliability				
Total Scores		.864		.817
Adaptive Scores		.796		.753

Subtest 5 on Circuit Measurement and Wave Form Analysis. As indicated in the total test reliability and validity indices, the impact on the adaptive testing paradigm of this hierarchical shift seemed to be minimal.

In regard to the interrelation indices, Tables 4 and 5 present the intercorrelation among performance and time variables. Most importantly, one should note that the correlation was essentially perfect between total scores and adaptive scores. As expected, there was a similar pattern for the Green scores. In looking at both the total score and adaptive score in relation to the subtests, one will notice that they were again similar in magnitude with very little perturbation within either of the test situations. On the other hand, the moderate intercorrelations among the five subtests would lead one to have questions about the accuracy of predicting a subsequent subtest on the basis of a lower one. This is undoubtedly constrained by the number of test items to be found in any given subtest. For example, Subtest 1 of Block IV, having 12 items, yielded a more consistent pattern. The intercorrelations among the higher level subtests also tended to be higher in magnitude, but not statistically different from zero.

The issue of validity is confounded in the current experiment in that the adaptive scores are a subset of the total scores. Path analysis (Kerlinger et al., 1973) offers a method for determining direct and indirect causal relationships. Using total scores as the dependent measure and the five subtest adapt scores as the predictor variables, the total direct effects (e.g., r_{15}) tended to be in the .5 to .7 range. The total indirect effects were in the .2 to .3 range, indicative of the hierarchical effects. While more detailed observations are possible, the path analysis outcomes evidenced a progressive causality relationship and the total indirect effects documented the hierarchical effects. Finally, the analysis established another form of the concurrent validity of adaptive scores to total scores.

There were nine students who participated in both the Block II

Table 4

Intercorrelations Among Performance and Time Variables for Block II Students (N = 133)

		1	2	3	4	5	6	7	8
1.	Total Score	-	.99	.98	.45	.57	.77	.67	.81
2.	Adapt Score		-	.96	.45	.60	.79	.68	.81
3.	Green Score			-	.43	.53	.68	.61	.77
4.	Test 1 Score				-	.27	.23	.25	.36
5.	Test 2 Score					-	.44	.38	.42
6.	Test 3 Score						-	.43	.68
7.	Test 4 Score							-	.45
8.	Test 5 Score								-

Table 5

Intercorrelations Among Performance and Time Variables for Block IV Students (N = 133)

		1	2	3	4	5	6	7	8
1.	Total Score	-	.99	.90	.79	.56	.62	.62	.53
2.	Adapt Score		-	89	.78	.57	.63	.63	.51
3.	Green Score			-	.75	.54	.63	.60	.65
4.	Test 1 Score				-	.24	.42	.42	.47
5.	Test 2 Score					-	28	.23	.19
6.	Test 3 Score						-	.40	.39
7.	Test 4 Score							-	.21
8.	Test 5 Score								-

and Block IV adaptive testing. A review of their means indicated that as a group they were approximately one-half standard deviation higher than the total groups. In regard to behavioral stability across unit tests, the correlations on the total score and adaptive score were in the low 60's (the correlation coefficient for total score was $r = .632$, $p < .05$; and the total adaptive score correlation coefficient was $r = .641$, $p < .05$). The remainder of the performance and time variables tended to be in the low-moderate range (.30 to .45). This added further support to the consistency of the adaptive instructional paradigm across the two testing situations,

Discussion

The primary focus of this study was concerned with generalizing an adaptive testing paradigm to a hierarchical conceptual course situation. From a validity point of view, a direct comparison of the adaptive test scores with the total test scores yielded a nearly perfect correlation (Block II--$r = .99$ and Block IV--$r = .99$). The path analysis procedures supported this causal relationship. The mean values and standard deviations for both performance and time variables were approximately the same magnitude. In an earlier study within a conventional Air Force technical training course it was found that the adaptive test scores correlated with total test scores at a highly significant level ($r = .940$) (Hansen et al., 1976). Thus the adaptive testing paradigm gained further support concerning its generalizability as a method for making instuctional decisions.

In reference to the generalizability of the testing time reduction, Blocks II and IV had an approximately 23 to 29 percent reduction in items with associated time reduction of 25 to 30 percent. Prior studies (Waters, 1975) have tended to find item reductions of up to 50 percent and time reductions of approximately 40 percent. The prior adaptive testing study on technical training in the Inventory Management area (Hansen et al., 1976) found that the adaptive testing paradigm yielded a 39.5 percent time reduction, remarkably consistent with Waters' result (1975). These reductions for adaptive testing seemed to have a consistent magnitude, namely, the more complex problem-solving oriented the items, the less likely that one will achieve a 50 percent reduction in the item or time savings. This was understandable when one considered that each of the 40 items required significant amounts of mental processing time and applications of rules in order to find the correct solution. While the questions were posed as multiple-choice in nature, in fact, they had to be worked out in a paper-and-pencil process.

This interpretation was further supported by noting that the item processing times were fairly consistent, about 1.6 minutes per item. Most significantly, the times per item for the adaptive section were slightly, but consistently, less than that for the post cutoff items; six out of ten comparisons favored the adaptive items. This phenomenon can be explained in two fashions. First, each of the items required considerable, but equivalent, mental processing time in order to execute all of the solution steps. The item time variability across students was considerably greater than between items. The correlations of adaptive time to adaptive scores were $r = -.48$ and $R = -.52$; this indicated the consistent student variability while means and standard deviations were approximately of the same magnitude. More importantly, though, those students who performed in the lower quarter of the performance range were likely to exit at the easier end of the test item sequence and consequently to have to encounter more difficult items. This was substantiated by noting that less than 30 percent of the items formed the post-cutoff pool; and approximately 37 percent of the students contributed over 90 percent of the responses to this pool; that is, 37 percent of the students exited at the easy end of any given subtest. Stated more simply, the better students were prone to always exit from the hardest end of the subtest and therefore to have higher performance and faster item times in the post-item range. Approximately, 37 percent of the students were likely to consistently exit from the easiest end and to have to encounter difficult items which were observed to have excessively long question /answer latencies. Many of these individual item times were three to four times as long compared to those yielded by the better performing students. These observations further substantiate the view that item testing time will have to be conditionalized by strata of performance.

In regard to test reliability, it was found that the adaptive test, even when composed of few items, yielded indices that were nearly equivalent to that of the total test. This outcome was equivalent to the prior adaptive testing study (Hansen et al., 1976) in technical training. It should be noted that the reliability indices as well as validity indices were higher in magnitude in this study.

The hierarchical relationships tended to be revealed in the subtest mean item difficulties and the adaptive score by subtest correlations. The correlations among the subtest means yielded a low to moderate range of coefficients but substantially heightened during path analysis. There was a tendency for performance to diminish as one moved from the easier beginning subtest to the more difficult final subtest. On the other hand, there were decided

reversals, especially in the final subtest, which were most unusual in hierarchical learning situations. Perhaps this could be explained by the intensive learning offered during the two weeks of each of the blocks. The student-to-instructor ratio was approximately ten to one and the use of within-class testing as well as performance assessments undoubtedly provided extensive acquisition of most of the hierarchical structure. Therefore, one would expect minor reversals and a less clear hierarchical pattern in situations where highly effective training was forthcoming. These reversal shifts might imply that a total adaptive approach (ranking all items by difficulty) might be more effective in contrast with the task-analyzed subtests as reflected in this study. These contrasting alternatives have to be weighted in regard to the diagnostic benefit of a student's performance on a given subtest conceptual set as opposed to further reductions in testing items.

As indicated, nine students participated in both Block II and Block IV adaptive tests. The outcomes indicated that a moderate level of unit-to-unit consistency was found. (The adaptive test correlations were .63 and .64, respectively.) These results added further evidences as to the consistency of the adaptive testing paradigm.

Adaptive testing has been applied in ability assessment situations, in lower level technical training, and now in highly complex hieratchically structured technical training. In most respects the results have been quite consistent; namely, adaptive testing yielded equivalent reliability and validity indices while reducing the number of testing items and the overall testing time. The amount of reduction appeared to be a direct function of the amount of complexity both in the course material as well as the test items. The empirical outcomes for adaptive testing have now been sufficiently conclusive that one can look forward to their operational application within the near future.

References

Gagne, R.M. The acquisition of knowledge. Psychological Review 69 (July 1962): 355-365.

Green, B.G. Comments on tailored testing. In W. Holzman (Ed.), Computer-Assisted Instruction, Testing, and Guidance. New York: Harper and Row, 1970.

Haladyna, T.M. An Investigation of Full and Subscale Reliabilities of Criterion-Referenced Tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974. ED 091 435.

Hansen, D.N., Harris, R., & Ross, S. Flexilevel Adaptive Testing Paradigm: Validation in Technical Training. AFHRL-TR-77-35 (I). Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, July 1977.

Kerlinger, F.N., & Pedhazur, E.J. Multiple Regression In Behavioral Research. New York: Holt, Rinehart and Winston, 1973.

Larkin, K.C., & Weiss, D.J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report, 75-1, University of Minnesota, Minneapolis, 1975.

Lord, F.M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 31 (Winter 1971): 805-813

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill, 1967.

Nunnally, J.C., & Durham, R.L. Validity, reliability and special problems of measurement in evaluation research. In E.L. Struening & M. Guttentag, (Eds.), Handbook of Evaluation Research (Vol. 1). Beverly Hills: Sage, 1975.

Waters, B.K. Empirical investigation of the stradaptive testing model for the measurement of human ability. AFHRL, TR-75-27, AD-A018611. Williams AFB, AZ; Flying Training Division, Air Force Human Resources Laboratory, October 1975.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1977 – 771-057/45